Anatomical Priors in Convolutional Networks for Unsupervised Biomedical Segmentation

Adrian V. Dalca
MIT and MGH
adalca@mit.edu

John Guttag
MIT
guttag@mit.edu

Mert R. Sabuncu
Cornell University
msabuncu@cornell.edu

Abstract

We consider the problem of segmenting a biomedical image into anatomical regions of interest. We specifically address the frequent scenario where we have no paired training data that contains images and their manual segmentations. Instead, we employ unpaired segmentation images that we use to build an anatomical prior. Critically these segmentations can be derived from imaging data from a different dataset and imaging modality than the current task. We introduce a generative probabilistic model that employs the learned prior through a convolutional neural network to compute segmentations in an unsupervised setting. We conducted an empirical analysis of the proposed approach in the context of structural brain MRI segmentation, using a multi-study dataset of more than 14,000 scans. Our results show that an anatomical prior enables fast unsupervised segmentation which is typically not possible using standard convolutional networks. The integration of anatomical priors can facilitate CNN-based anatomical segmentation in a range of novel clinical problems, where few or no annotations are available and thus standard networks are not trainable. The code, model definitions and model weights are freely available at http://github.com/adalca/neuron.

1. Introduction

Biomedical image segmentation plays a crucial role in many applications, such as population analysis, disease progression modelling, or treatment planning. Convolutional neural networks (CNNs), a class of deep learning methods, have recently been employed to derive powerful biomedical segmentation algorithms, showing promise of overcoming limitations in previous methods [3, 4, 29, 34]. However, CNN-based approaches most often depend on (large-scale) training data, particularly in the form of image scans paired with segmentations. These annotations are often costly and challenging to obtain because they require the tedious effort of a trained expert, taking several expert hours per scan.

1.1. Contributions

To our knowledge, there has not been a theoretically rigorous effort to integrate rich probabilistic anatomical priors with a CNN-based segmentation model in a computationally effective manner. We introduce a generative model for biomedical segmentation that employs an anatomical prior. We describe a principled theoretical derivation that follows directly from our generative model. We demonstrate that this yields intuitive cost functions and simpler models. We use an auto-encoding variational CNN to characterize the anatomical prior, and an encoder-decoder CNN to provide fast segmentation of medical images in unsupervised settings.

We demonstrate the method in an unsupervised biomedical image segmentation setting where paired annotations are not available. Our proposed strategy is general and computationally efficient, provides a natural framework for sampling possible subject-specific segmentations of a scan, and provides uncertainty estimates for these segmentations.

2. Related Work

2.1. Segmentation Convolutional Neural Networks

CNN-based segmentation approaches generally rely on fully convolutional architectures applied to image data. They extract hierarchical and multi-resolution features that are in turn combined to compute a semantic segmentation [23, 29, 31, 34].

A popular discriminative segmentation architecture, U-net [29], involves a convolutional encoder or downsampling network, followed by a convolutional decoder or upsampling network, and skip-connections between layers. The encoder captures relevant features of the input image at different resolutions. The decoder then synthesizes a high-resolution segmentation, using the skip connections to achieve voxel-level precision. While the exact architecture of these networks, such as the number of layers and levels, size of convolution kernels, or application of batch normal-

ization vary, they typically involve millions of parameters and necessitate large datasets and data augmentation techniques to train.

CNN-based segmentation models have two major shortcomings: the dependency on annotated data, limiting their use in unsupervised settings; and their lack of anatomical knowledge. The latter limits the network's ability to be faithful to known anatomical shapes during segmentation.

In our work, we use CNN architectures to learn anatomical priors and segment medical images. The prior eliminates the burden of providing paired example segmentations.

2.2. Priors for Convolutional Neural Networks

A clinical expert performing manual delineation relies on spatial coordinates and prior knowledge about anatomy, and may use a template of the structures to constrain the task. This process draws on the anatomical similarity across patient scans. This is in stark contrast with typical computer vision problems that have led to many popular CNN architectures, where object location, shape, and appearance can be unpredictable.

Convolutional methods are often limited in incorporating domain expertise. For example, U-Net [29] and its derivatives produce segmentation algorithms that do not exploit location information or other explicit anatomical priors. A CNN might have difficulty differentiating two distinct objects that are consistently in two specific parts of the scan, if they have the same intensity and context (as in bilateral structures in two hemispheres)[1]. While increasingly more complex networks that extend receptive fields may tease out object differences in supervised settings, the problem would be trivial if we consider anatomical knowledge like spatial location. Furthermore, in these modalities, image contrast can be weak or noisy in certain regions resulting in uncertainty of the segmentations. An anatomical prior can resolve these ambiguities, while making the segmentation task easier.

A popular strategy to explicitly employ prior structure in CNNs for biomedical image segmentation is to use a conditional random field (CRF) as a post-processing step [13, 30, 34]. However, CRFs only captures local constraints, and adds to the computational burden. Location information has been included as a feature in patch-based CNN segmentation networks [34]. While this addition carries prior location information, it is network-specific, increases the parameter burden on the network, and does not capture shape information.

Recent methods have employed shape priors for neural network solutions in supervised problems [26, 28]. In particular, they often design a series of networks that learn representations of images and segmentations in a supervised setting. They propose *ad-hoc* cost functions that encourage the computed segmentations to be similar to both the learned shape and the ground truth. These methods attempt to correct segmentations produced by standard CNNs by adding a prior constraint.

Convolutional image generative models, such as generative adversarial nets, have grown in popularity. They have recently been applied to biomedical image segmentations [16, 24] in a supervised setting where standard loss functions are combined with adversarial losses. A series of recent papers in the computer vision community removes the requirement for paired data by introducing a cycle dependency [37]. However, these methods are less applicable in medical image segmentation with many anatomical labels, as an image signal can pass through the rich networks at low cost, leading to a perfect cycle loss, circumventing the required constraints [37].

Variational Bayes auto-encoders have been used for various tasks to learn probabilistic generative models, and often use convolutional networks [18]. Our method builds on these models to combine anatomical priors with image generation.

2.3. Classical Generative Models

Encoding and exploiting prior knowledge is common in generative models. Our inspiration comes from classical atlas-based probabilistic segmentation methods that estimate the maximum *a posteriori* (MAP) probability based on a generative model involving a prior probability and likelihood [5, 10, 17, 27, 32, 33, 35].

The prior term captures knowledge of underlying anatomy and usually involves a probabilistic atlas and a spatial deformation that models geometric variation. The spatial deformation can be explicitly solved using a registration algorithm or accounted for in a unified segmentation framework [1].

The likelihood models the physical process that yields medical image intensities, sometimes called the appearance model, conditioned on the latent anatomy. These appearance models are often simpler, relying on additive and/or multiplicative Gaussian or Rician noise models [35]. Model parameters are most often estimated using training data, such as annotated image pairs, for example using maximum likelihood.

Given a new image, most popular segmentation algorithms use numerical non-convex optimization and can take several hours per image on a modern CPU.

In our model, we draw on ideas from classical model-based biomedical segmentation algorithms, convolutional neural networks (CNNs) used in semantic segmentation, and recent developments in variational Bayes approximations using neural networks. In our experiments, we con-

[1] Assuming that the field of view of the network is constrained to not include the other object's vicinity

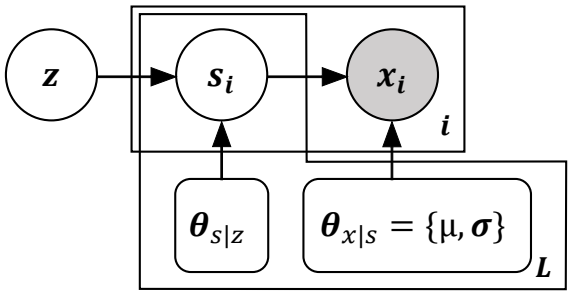


Figure 1. A graphical representation of our generative model. Circles indicate random variables and rounded squares represent parameters. Shaded circles represent observed quantities and the plates indicate replication. $\boldsymbol{x}_i$ is the acquired image. The image intensities are generated from a normal distribution parametrized by $\boldsymbol{\mu}_l$ and σ_l for each anatomical label l in the label map $\boldsymbol{s}$. Anatomical priors are controlled by the variable $\boldsymbol{z}$ and categorical parameters $\boldsymbol{\theta}_{s|z}$.

sider the segmentation of structural brain MRI scans into cortical and subcortical regions of interest (ROIs). Our results show that the proposed anatomical prior enables rapid unsupervised segmentation. While complex, specialized tools exist for segmenting some specific scan modalities or particular diseases, they do not generalize to other modalities and can take hours to process one scan. Our goal is to provide a first general approach to biomedical image segmentation in an unsupervised setting.

3. Generative Model

We let $\boldsymbol{x}$ be an (MR) 3D volume, and assume it is generated from a 3D anatomical segmentation map $\boldsymbol{s}$. We will use $x[j]$ and $y[j]$ to denote the image intensity and label at voxel j, respectively.

We use a generative model to describe the spatial distribution, shape, and appearance of anatomical structures. Figure 1 provides a graphical representation.

The prior captures our knowledge about spatial distributions and shape of anatomy. We let $\boldsymbol{z}$ be a latent variable representing an embedding of these shapes, and model the prior probability of this embedding as normal with mean $\mathbf{0}$ and an identity covariance matrix:

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \mathbb{1}), \tag{1}$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the normal distribution parametrized by mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

We let $\boldsymbol{s}$ be drawn from a categorical prior distribution determined by the low-dimensional embedding $\boldsymbol{z}$ via $p_{\boldsymbol{\theta}_{s|z}}(\boldsymbol{s}|\boldsymbol{z})$:

$$p_{\boldsymbol{\theta}_{s|z}}(\boldsymbol{s}|\boldsymbol{z}) = \prod_j f_{j,s[j]}(\boldsymbol{z}|\boldsymbol{\theta}_{s|z}) \tag{2}$$

where $f_{j,l}(\cdot; \boldsymbol{\theta}_{s|z})$ is the probability of label l at voxel j.

Finally, given the label map $\boldsymbol{s}$, the intensity observations are generated via $p_{\boldsymbol{\theta}_{x|s}}(\boldsymbol{x}|\boldsymbol{s})$, sampled at each voxel from a normal distribution:

$$p_{\boldsymbol{\theta}_{x|s}}(\boldsymbol{x}|\boldsymbol{s}) = \prod_j \prod_l \mathcal{N}(\boldsymbol{x}[j]; \boldsymbol{\mu}_l, \sigma_l)^{\delta(\boldsymbol{s}[j]=l)}, \tag{3}$$

where $\boldsymbol{\theta}_{z|s} = \{\boldsymbol{\mu}_l, \sigma_l\}$, and $\delta(\boldsymbol{s}[j] = l)$ is the indicator function that evaluates to 1 if $\boldsymbol{s}[j] = l$ and 0 otherwise. The joint likelihood is therefore $p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{s}|\boldsymbol{z}) = p_{\boldsymbol{\theta}_{x|s}}(\boldsymbol{x}|\boldsymbol{s}) p_{\boldsymbol{\theta}_{s|z}}(\boldsymbol{s}|\boldsymbol{z})$, where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{s|z}, \boldsymbol{\theta}_{x|s}\}$.

We describe the learning procedure in the next section. Given learned parameters, to obtain the segmentation $\boldsymbol{s}_i$ given a new image $\boldsymbol{x}_i$, we perform MAP estimation:

$$\begin{aligned} \hat{\boldsymbol{s}}_i &= \arg\max_{\boldsymbol{s}_i} \log p(\boldsymbol{s}_i|\boldsymbol{x}_i; \boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{s}_i} \log p(\boldsymbol{s}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) \end{aligned} \tag{4}$$

4. Learning

In this section, we describe a learning strategy that uses convolutional neural networks to estimate anatomical representations and optimize posterior segmentation distributions. This procedure is applicable to broad modelling choices for the probability distributions described above. We also discuss a separate learning procedure for the anatomical prior, uncertainty estimation, and implementation.

Without assuming voxel independence of the segmentation map given an image, estimating the posterior probability $p_{\boldsymbol{\theta}}(\boldsymbol{s}|\boldsymbol{x})$ is intractable since it involves integrating over the latent variable $\boldsymbol{z}$. Estimating $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s})$ is similarly intractable, making the Expectation Maximization algorithm not pertinent.

We first introduce an encoding probability $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s})$ as an approximation to the intractable $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s})$, similar to [18]. Consider the KL divergence between the approximate distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s})$ and the true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s})$:

$$\begin{aligned} &\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s}) || p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s})\right] \\ &\quad = \mathbf{E}_q\left[\log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s}) - \log p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s})\right] \\ &\quad = \mathbf{E}_q\left[\log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s}) - \log p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{z})\right] + \log p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{s}). \end{aligned} \tag{5}$$

Rearranging terms, we obtain

$$\begin{aligned} \log p(\boldsymbol{x}, \boldsymbol{s}) = &\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s}) || p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s})\right] \\ &+ \mathbf{E}_q\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{z}) - \log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{s})\right]. \end{aligned} \tag{6}$$

Since the KL divergence of the approximate and true posterior of $\boldsymbol{z}$ is non-negative, the second term is referred to as the *variational lower bound* of the model evidence

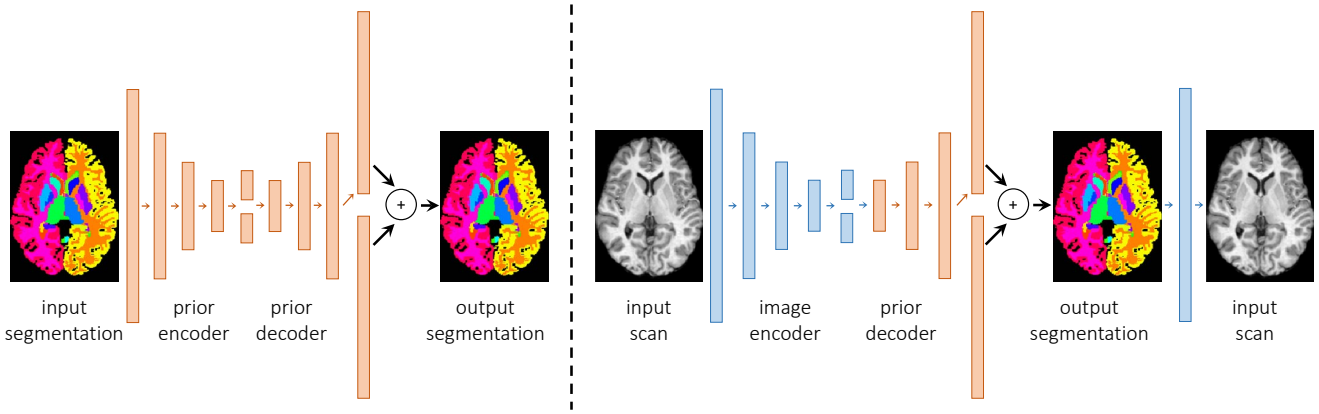


Figure 2. Left: **Proposed Auto-Encoding Variational Anatomical Prior**. A variational auto-encoder is combined with a location-specific prior layer. Right: **Proposed architecture for learning generative model parameters**. Orange and blue arrows indicate down/up-sampling in the prior and full model, respectively, and rectangles represent a stack of convolutional layers with non-linearities, with their heights reflecting the size of the vectors.

or joint probability. For a given approximate distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s})$, we can estimate $\boldsymbol{\theta}$ by optimizing the lower bound:

$$\begin{aligned}\mathcal{V}_{\text{model}}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x},\boldsymbol{s}) \\ &= \mathbf{E}_q\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{s},\boldsymbol{z}) - \log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s})\right] \\ &= \mathbf{E}_q\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{s}|\boldsymbol{z})\right] - \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s})||p(\boldsymbol{z})\right].\end{aligned} \tag{7}$$

We model the approximating posterior $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s})$ as a normal that depends on the image only:

$$\begin{aligned}q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{s}) &= q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \\ &= \mathcal{N}(\boldsymbol{z};\boldsymbol{\mu}_{z|x},\boldsymbol{\Sigma}_{z|x}).\end{aligned} \tag{8}$$

where $\boldsymbol{\Sigma}_{z|x}$ is diagonal.

We estimate the parameters of the approximating distribution using convolutional neural networks. We design an encoding convolutional neural network $\text{enc}_{\boldsymbol{\phi}}(\boldsymbol{x})$ that takes as input $\boldsymbol{x}$ and outputs the parameters of the approximating posterior distribution $\boldsymbol{\mu}_{z|x}(\boldsymbol{x})$, and $\boldsymbol{\Sigma}_{z|x}(\boldsymbol{x})$. This network learns how to embed an entire (MR) image into the most likely low-dimensional anatomical embedding $\boldsymbol{z}$ and its variance.

Conditioned on $\boldsymbol{z}$, the probability of the segmentation can be computed with a *decoder* network $\text{dec}_{\boldsymbol{\theta}_{s|z}}(\boldsymbol{z})$ that takes $\boldsymbol{z}$ as input and outputs the parameters $f(\boldsymbol{z};\boldsymbol{\theta}_{s|z})$ of the segmentation categorical distribution $p_{\boldsymbol{\theta}_{s|z}}(\boldsymbol{s}|\boldsymbol{z})$. The parameters $\boldsymbol{\theta}_{s|z}$ of this decoder can be learned using a separate set of segmentations, as described below.

The final part of the generative model, the appearance or likelihood model, can also be learned with a neural network that takes a segmentation probability map as input and computes the parameters $\boldsymbol{\mu}_l$. We separately estimate σ_l, assuming additive zero mean Gaussian noise in an image, using a difference of Laplacian filters [15].

4.1. Auto-Encoding Anatomical Prior

In this work, we learn a prior independently from an unpaired segmentation dataset. This enables the flexibility of having an external description of the anatomy that need not be available in the current data. Unfortunately, as before, estimating the probability distribution $p(\boldsymbol{s})$ is intractable. Following a derivation similar to the previous section and to the auto-encoding variational Bayes framework, we introduce an approximation $q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{s})$ to the posterior $p(\boldsymbol{z}|\boldsymbol{s})$ as a normal distribution:

$$q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{s}) = \mathcal{N}(\boldsymbol{z};\boldsymbol{\mu}_{z|s},\boldsymbol{\Sigma}_{z|s}), \tag{9}$$

where $\boldsymbol{\Sigma}_{z|s}$ is diagonal, leading to the following lower bound:

$$\begin{aligned}\mathcal{V}_{\text{prior}}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{s}) \\ &= \mathbf{E}_q\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{s},\boldsymbol{z}) - \log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{s})\right] \\ &= \mathbf{E}_q\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{s}|\boldsymbol{z})\right] - \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{s})||p(\boldsymbol{z})\right].\end{aligned} \tag{10}$$

This optimization can be solved using a Stochastic Gradient Variational Bayes (SGVB) estimator that uses mini-batches. The reparametrization trick allows us to sample $\boldsymbol{z}_k \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{s})$, leading to an approximation of the expectation $\mathbf{E}[\cdot]$ [18]. The loss $\mathcal{L}_i(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{s}_i)$ for each data point $\boldsymbol{s}_i$ and sample $\boldsymbol{z}_k \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{s}_i)$ is

$$\begin{aligned}\mathcal{L}_{\text{prior}}(\boldsymbol{\theta}_{s|z},\boldsymbol{\psi};\boldsymbol{s}_i,\boldsymbol{z}_k) \\ &= \mathrm{KL}\left[\log q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{s}_i)||\log p(\boldsymbol{z})\right] - \log p_{\boldsymbol{\theta}_{s|z}}(\boldsymbol{s}_i|\boldsymbol{z}_k) \\ &= \frac{1}{2}\sum_j (1 + \log\left(\boldsymbol{\Sigma}_{z|s_i}[j]\right) - \boldsymbol{\mu}^2_{z|s_i}[j] - \boldsymbol{\Sigma}_{z|s_i}[j]\Big) \\ &\quad - \sum_j \boldsymbol{s}_i[j]\log f(\boldsymbol{z}_k;\boldsymbol{\theta}_{s|z})[j].\end{aligned} \tag{11}$$

We design an encoding network $\text{enc}_{\psi}(\boldsymbol{s})$ that takes a segmentation map as input and outputs the parameters $\boldsymbol{\mu}_{z|s}$ and $\boldsymbol{\Sigma}_{z|s}$. Importantly, we learn the parameters of the encoding network given only a set of segmentations $\{\boldsymbol{s}_i\}$, which can be derived from other imaging modalities and/or datasets. The segmentation prior therefore does not require paired training data in the traditional sense. For example, we can use a prior computed using publicly available annotated datasets such as [19] in a problem that involves a different imaging modality than in the current task.

4.2. Unsupervised Learning

We assume we have learned a segmentation prior using the Auto-Encoding Anatomical Prior described in the previous section. In particular, we will utilize the *decoder* component of the prior model, namely $p_{\boldsymbol{\theta}_{s|z}}(\boldsymbol{s}|\boldsymbol{z})$.

If we had annotated pairs $\{\boldsymbol{x}_i, \boldsymbol{s}_i\}$, we could jointly learn model parameters $\boldsymbol{\theta}_{x|s}$, and variational parameters $\boldsymbol{\phi}$ by optimizing the evidence lower bound objective, similar to the previous section. For each sample $\{\boldsymbol{x}_i, \boldsymbol{s}_i\}$ and sample $\boldsymbol{z}_k \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_i, \boldsymbol{s}_i)$, the loss function would be

$$
\begin{aligned}
&\mathcal{L}_{\text{model}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_i, \boldsymbol{s}_i, \boldsymbol{z}_k) = -\mathcal{V}_i(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}_i, \boldsymbol{s}_i, \boldsymbol{z}_k) \\
&\quad = \text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_i)||p(\boldsymbol{z})\right] - \log p_{\boldsymbol{\theta}_{s|z}}(\boldsymbol{x}_i, \boldsymbol{s}_i|\boldsymbol{z}_k). \\
&\quad = \text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_i)||p(\boldsymbol{z})\right] - \log p_{\boldsymbol{\theta}_{s|z}}(\boldsymbol{s}_i|\boldsymbol{z}_k) \\
&\qquad - \log p_{\boldsymbol{\theta}_{x|s}}(\boldsymbol{x}_i|\boldsymbol{s}_i). \\
&\quad = \frac{1}{2}\sum_j (1 + \log\left(\boldsymbol{\Sigma}_{z|x_i}[j]\right) - \boldsymbol{\mu}^2_{z|x_i}[j] - \boldsymbol{\Sigma}_{z|x_i}[j]\right) \\
&\qquad - \sum_j \boldsymbol{s}_i[j] \log f(\boldsymbol{z}; \boldsymbol{\theta}_{s|z})[j] \qquad (12) \\
&\qquad + \sum_j \sum_l \frac{\delta(\boldsymbol{s}_i[j] = l)}{2\sigma_l^2}(\boldsymbol{x}_i - \boldsymbol{\mu}_l).
\end{aligned}
$$

resulting in terms of KL divergence, segmentation map categorical cross-entropy, and intensity-based mean squared error, respectively. During training, these terms would ensure that the probability $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_i)$ stays close to the standard normal, while explaining the segmentations, and that the model parameters $\boldsymbol{\theta}_{x|s} = \{\boldsymbol{\mu}_l, \boldsymbol{\sigma}_l\}$ capture the relationship between the segmentations and the images.

However, in this paper we tackle the unsupervised setting, where annotated pairs $\{\boldsymbol{x}_i, \boldsymbol{s}_i\}$ are not available, and we only have the images $\{\boldsymbol{x}_i\}$. Therefore, we cannot compute the categorical cross entropy term in (12). Instead, we marginalize over the segmentation $\boldsymbol{s}$ in the second term of the variational lower bound (7):

$$
\begin{aligned}
&\mathbf{E}_q\left[\log \int_{\boldsymbol{s}} p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{s}|\boldsymbol{z}) d\boldsymbol{s}\right] \\
&\quad = \mathbf{E}_q\left[\log \int_{\boldsymbol{s}} p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{s}) p_{\boldsymbol{\theta}_{s|z}}(\boldsymbol{s}|\boldsymbol{z}) d\boldsymbol{s}\right] \\
&\quad \geq \mathbf{E}_q\left[\int_{\boldsymbol{s}} p_{\boldsymbol{\theta}_{s|z}}(\boldsymbol{s}|\boldsymbol{z}) \log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{s}) d\boldsymbol{s}\right] \\
&\quad = \mathbf{E}_{q, p(\boldsymbol{s}|\boldsymbol{z})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{s})\right] \qquad (13)
\end{aligned}
$$

where we used Jensen's inequality. We therefore arrive at the following upper bound of the loss function:

$$
\begin{aligned}
&\mathcal{L}_{\text{model}}(\boldsymbol{\theta}_{x|s}, \boldsymbol{\phi}; \boldsymbol{x}_i, \boldsymbol{z}_i) \\
&\quad = \frac{1}{2}\sum_j (1 + \log\left(\boldsymbol{\Sigma}_{z|x_i}[j]\right) - \boldsymbol{\mu}^2_{z|x_i}[j] - \boldsymbol{\Sigma}_{z|x_i}[j]\right) \\
&\qquad + \sum_j \sum_l \frac{f_{j,l}(\boldsymbol{z}_k|\theta_{s|z})}{2\sigma_l^2}(\boldsymbol{x}_i - \boldsymbol{\mu}_l), \qquad (14)
\end{aligned}
$$

where we used the factorization of $p_{\theta s|z}(\boldsymbol{s}|\boldsymbol{z})$ over voxels from (2), and sample $\boldsymbol{z}_k \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}_i)$.

4.3. Inference and uncertainty

Given a new image $\boldsymbol{x}$, we compute $\widehat{\boldsymbol{s}} = \arg\max_{\boldsymbol{s}} p_{\boldsymbol{\theta}}(\boldsymbol{s}|\boldsymbol{x})$ by first obtaining $\boldsymbol{\mu}_z$ using the encoder $\text{enc}_{\boldsymbol{\theta}_{z|x}}(\boldsymbol{x})$, and taking the maximum segmentation at each voxel $\widehat{\boldsymbol{s}} = \arg\max_{\boldsymbol{s}} \text{dec}_{\boldsymbol{\theta}_{s|z}}(\boldsymbol{\mu}_z)$. The operations are fast, since both are feed-forward neural networks.

This model also enables sampling segmentations conditioned on a particular image and enables estimation of uncertainty. Given an input image $\boldsymbol{x}_i$, we can create samples $\boldsymbol{z}_k \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and $\boldsymbol{s}_k \sim p_{\boldsymbol{\theta}_{s|z}}(\boldsymbol{s}|\boldsymbol{z}_k)$, simulating different plausible segmentations for a given subject. We can estimate the uncertainty of our segmentation given a new image $\boldsymbol{x}_i$ using

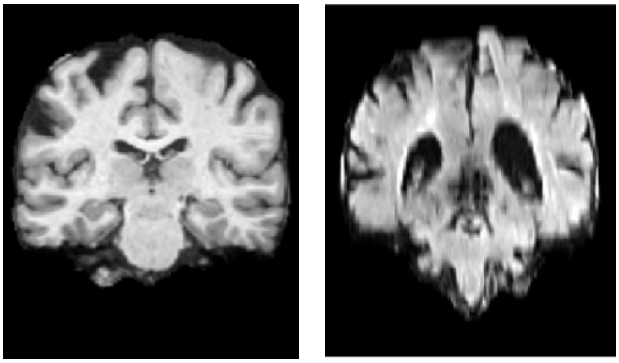
$$
\begin{aligned}
H(\boldsymbol{s}[j]) &= \mathbf{E}[-\log(p(\boldsymbol{s}[j]|\boldsymbol{x}_i))] \qquad (15) \\
&= -\sum_l p(\delta(\boldsymbol{s}[j] = l)|\boldsymbol{x}_i) \log(p(\delta(\boldsymbol{s}[j] = l)|\boldsymbol{x}_i)).
\end{aligned}
$$

4.4. Implementation

A CNN can be seen as a hierarchical function, a set of concatenated functions, or layers. For example, CNNs often map some input image $\boldsymbol{x}$ to an output probability $\hat{\boldsymbol{s}}_p$:

$$
\hat{\boldsymbol{s}}_p = \boldsymbol{f}_L \circ \ldots \circ \boldsymbol{f}_1(\boldsymbol{x}), \qquad (16)
$$

where $\circ$ denotes concatenation, $\boldsymbol{f}_i$ is often some nonlinear function such as a rectified linear unit or ReLU or max pool [12] applied to (linear) convolutions of the output of the previous layer $\boldsymbol{f}_{i-1}$ (with $\boldsymbol{f}_0 = \boldsymbol{x}$).



T1w scan T2-FLAIR scan

Figure 3. **Example T1w and T2-FLAIR images** highlighting the difference in anatomical differences, tissue contrast and scan quality.

Although we operate on 3D images, we use a 2D architecture in our experiments. We experimented with 3D architectures as well, but found little gain while facing significant challenges related to tradeoffs and limitations between GPU memory, batch size, the number of features, and the number of labels possible. Each encoder consists of five downsampling levels of one convolution layer each, with 3x3 convolution kernels with elu activations, and 32 features for each kernel. The final layer is dense, with 1000-long encoding of the means and standard deviations representations.

The decoder is a mirror of this design, but upsamples instead of downsampling and ends with a sigmoid activation. In addition, we use a final layer that implements a pixel-wise spatially-varying voxel-wise (location) prior $p_{\text{loc}}(\boldsymbol{s})$, which is multiplied with $\text{dec}(\boldsymbol{z})$ (in practice, we add the logarithms). As is common in the atlas-based segmentation literature, the prior $p_{\text{loc}}(\boldsymbol{s})$ was computed as the frequency of labels in the held out prior dataset, in affine-normalized coordinate system. This layer discourages any extreme decodings of $\boldsymbol{z}$ but does not capture shape properties, which is encoded in $\text{dec}(\boldsymbol{z})$.

We implement the normal probability $p(\boldsymbol{x}|\boldsymbol{s})$ with a single-layer linear network. We also find it useful to pre-train the image encoder using an image variational auto-encoder similar to the segmentation one. The encoder weights are used as initialization only. During training, we used the Adadelta optimizer [36].

For the latent encoding layers representing $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$, we introduce an activation function that discourages the sample activations from being too large, helping limit numerical issues stemming in sampling from these layers during the reparametrization trick. We use concepts from the softsign and tanh activations to define our function as $\text{act}(\boldsymbol{x}) = \text{softsign}_{\alpha}(\boldsymbol{x}) \log(2 + \alpha * |\boldsymbol{x}|)$.

5. Experiments

We demonstrate our model on two datasets. For the first dataset, we obtain ground truth segmentations using a specialized algorithm with intense computational requirements, combined with manual work and QC [9]. We use a subset of these segmentations to learn the prior probability parameters. We treat the rest of the dataset as *unsupervised*, where we only use the ground truth segmentations as validation. For the second dataset we do not have ground truth, offering a realistic scenario. Figure 3 shows example images from the two datasets, highlighting the difference, and the difficulty of the task.

5.1. Data

T1w scan dataset

We gathered a large-scale multi-site, multi-study dataset of more than 14,000 T1-weighted brain MRI scans from eight publicly available datasets: including data from ADNI [25], OASIS [20], ABIDE [7], ADHD200 [22], MCIC [11], PPMI [21], HABS [6], and Harvard GSP [14]. Subject age ranges, health states, and acquisition details vary with each dataset, but all scans were resampled to a 256x256x256 grid with 1mm isotropic voxels, and all images cropped to 160x192x224 to eliminate entirely-background voxels.

We carry out standard pre-processing steps, including affine spatial normalization using FreeSurfer for each scan [9]. All MRIs were also segmented with FreeSurfer - a task that takes several CPU hours per scan. We also applied quality control (QC) using visual inspection to catch gross errors in segmentation results.

We partitioned the data into a prior training subset of 5,000 images, where we only used the annotations. The rest of the data was treated as an unannotated dataset, where QCed segmentations were only used for validation.

While developing the network architectures we partitioned the rest of the data into training, validation and test sets. Once the architecture was fixed, we reported results on the test dataset by training and evaluating the model in an unsupervised fashion.

T2-FLAIR scan dataset

We also gathered a dataset of more than 3800 T2-FLAIR scans, a significantly different MR modality, from the ADNI cohort. These scans exhibit significantly different tissue properties compared to the T1w images, lower acquisition quality, and exhibit 5mm slice spacing (Figure 3). They provide a good test of our hypothesis that priors learned are useful for segmenting image data with different tissue properties. To our knowledge there is no automatic method to obtain detailed anatomical segmentations for these images. We affinely align these images to the same space as

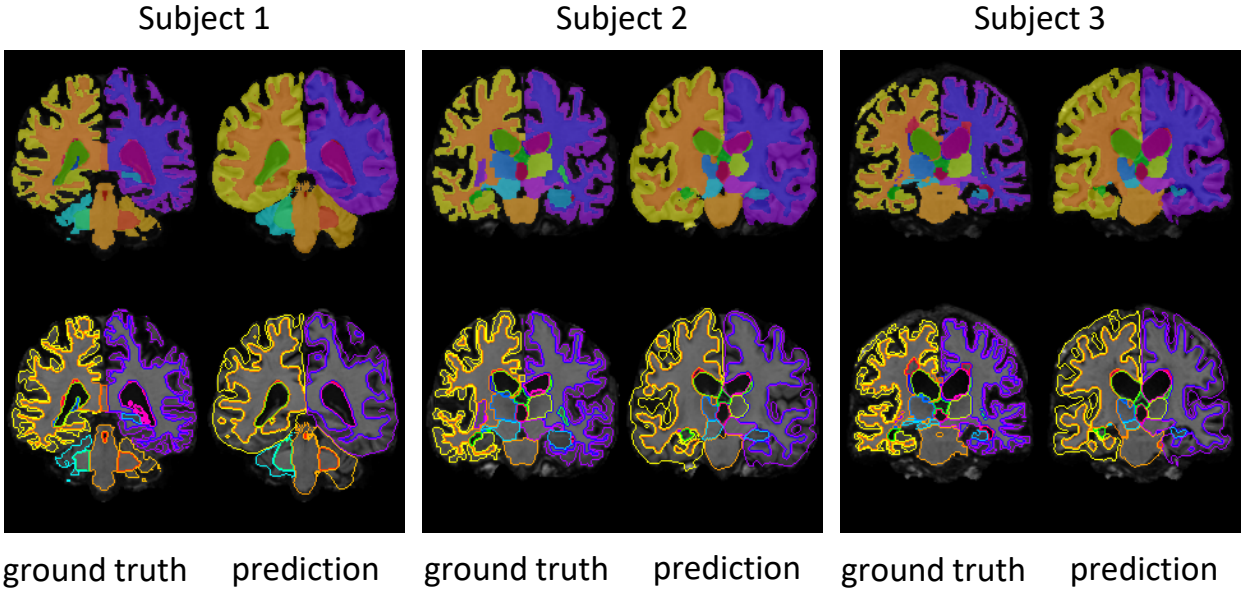


Figure 4. **T1 segmentation examples**. For each subject, the left column shows the "ground truth" as estimated with FreeSurfer, and the right illustrates our prediction. The first row overlays anatomical structures on top of the subject scan to clearly indicate the proposed segmentation. The second row shows outlines of each structure to allow comparison with the subject scan.

the T1 images using mutual information based registration with ANTs [2]. We perform brain extraction using an in-house developed neural network-based algorithm that uses a UNet architecture and extensive data augmentation.

In the set of annotations that we used to train the prior, we avoided including any annotations coming from subjects whose T2-FLAIR scans are in this dataset.

5.2. Evaluation

We evaluate our results both visually and quantitatively. For the **T1w dataset**, we use a volume overlap measure, Dice, to quantify the automatic segmentation results [8]:

$$\text{Dice}(\widehat{\boldsymbol{s}}, \boldsymbol{s}_t) = \frac{2\sum_j \delta(\widehat{\boldsymbol{s}}[j] = l)\delta(\boldsymbol{s}_t[j] = l)}{\sum_j \delta(\widehat{\boldsymbol{s}}[j] = l) + \sum_j \delta(\boldsymbol{s}_t[j] = l)}. \quad (17)$$

where $\widehat{\boldsymbol{s}}$ is the predicted segmentation map, and $\boldsymbol{s}_t$ indicates the ground truth (FreeSurfer) label at each location. A Dice score of 1 indicates perfect segmentation.

We experimented segmenting in the unsupervised setting with standard UNet architectures, using the image MSE and mutual information loss functions. Because of the many structures that share similar intensities, these architectures are not able to produce sensible segmentations that resemble the correct segmentations, and we omit them from these results. Classical unsupervised methods that include sophisticated prior anatomical information take a significant amount time to run, and for T1w we regard FreeSurfer results as an optimistic bound for the T1w data. However, as these methods tend to be focused on specific modalities, there is no annotation tool for cortical and subcortical regions in T2-FLAIR. We evaluate the T2-FLAIR segmentation visually in Figure 5.

5.3. Results

At test time, a new subject only needs to be affinely registered to a template, after which the proposed CNN model evaluates a segmentation estimate. The entire process takes less than a few seconds on an NVidia Titan X GPU.

Fig. 4 shows a series of example segmentations for the T1w dataset demonstrating that our method is able to estimate anatomical structures, reproducing the general location as well as the shape of structures. Fine details, such as details of cortex folding, is not easily captured by the prior encoding, leading to smooth segmentation predictions. Fig. 6 illustrates the average Dice measure across several anatomical regions for T1 scans. We focus on the most prevalent (larger) structures, which can also be evaluated in detail in the visualizations of Figure 4.

Fig. 5 demonstrates our algorithm on T2-FLAIR scans. Even with the significantly lower image quality and different tissue contrasts, our algorithm is able to produce visually sensible segmentations. Our method is able to utilize the prior information to predict plausible segmentations, even given challenging images in unsupervised scenarios.

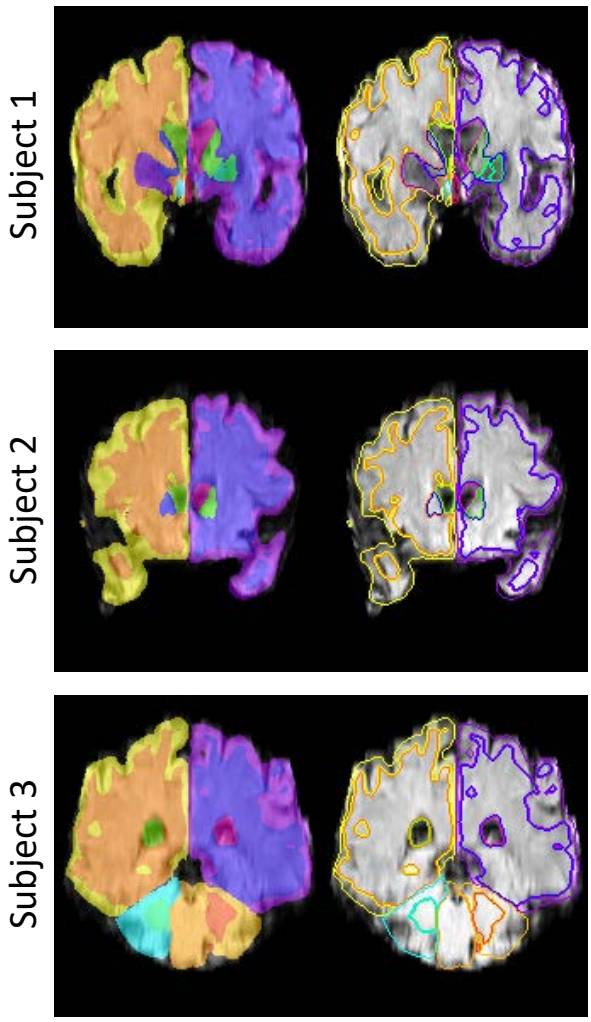


Figure 5. **T2-FLAIR segmentation examples**. The first column overlays anatomical structures on top of the subject scan to clearly indicate the proposed segmentation. The second column draws outlines of each structure to allow comparison with the subject scan. Different coronal slices illustrate the variability and difficulty of the task.

5.4. Discussion

Our method is able to reproduce anatomical structures that are guided by image contrast while respecting anatomical shapes according to the prior. Rapid, zero-shot segmentation is a challenging task, and to the best of our knowledge has not been tackled by previous methods. As such, the absence of prior results makes it difficult to fully interpret current results. The detailed FreeSurfer results are an upper bound, which any model is unlikely to achieve in the unsupervised setting. We omit showing results from lower bound (simplistic) baselines, such as the unsupervised U-Net model described above, since these models yielded non-

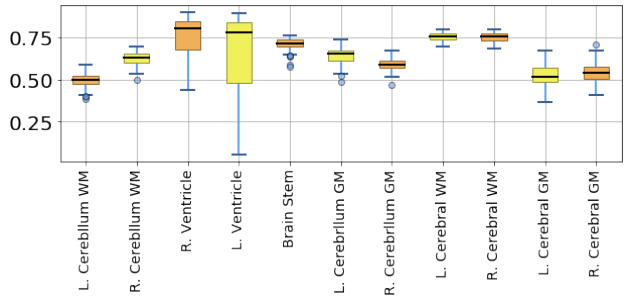


Figure 6. **Volume Overlap** measured via (Dice) for several structures in the T1w images.

sensical segmentations. To the best of our knowledge, our results are the first for zero-shot neural-network based segmentation of brain structures.

6. Conclusion

In this paper, we introduced a generative probabilistic model that employs a prior model learned through a convolutional neural network to compute segmentations in an unsupervised setting. We can interpret the anatomical prior as encouraging the neural network to predicting segmentation maps that come from a known distribution characterized by z while simultaneously producing images that agree with the observed scan. We demonstrate that our model enables segmentation using convolutional networks leading to rapid inference in a setting where segmentation is traditionally not possible, or takes hours to obtain for a single scan. The integration of priors promises to facilitate accurate anatomical segmentation in a variety of novel clinical problems with limited dataset availability.

References

[1] J. Ashburner and K. J. Friston. Unified segmentation. *Neuroimage*, 26(3):839–851, 2005. 2

[2] B. B. Avants, N. Tustison, and G. Song. Advanced normalization tools (ants). *Insight j*, 2:1–35, 2009. 7

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015. 1

[4] H. Chen, X. J. Qi, J. Z. Cheng, and P. A. Heng. Deep contextual networks for neuronal structure segmentation. In *Thirtieth AAAI conference on artificial intelligence*, 2016. 1

[5] O. Colliot, O. Camara, and I. Bloch. Integration of fuzzy spatial relations in deformable modelsapplication to brain MRI segmentation. *Pattern recognition*, 39(8):1401–1414, 2006. 2

[6] A. Dagley, M. LaPoint, W. Huijbers, T. Hedden, D. G. McLaren, J. P. Chatwal, K. V. Papp, R. E. Amariglio, D. Blacker, D. M. Rentz, et al. Harvard aging brain study: dataset and accessibility. *NeuroImage*, 2015. 6

[7] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer,

M. Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014. 6

[8] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 7

[9] B. Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012. 6

[10] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. Van Der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002. 2

[11] R. L. Gollub, J. M. Shoemaker, M. D. King, T. White, S. Ehrlich, S. R. Sponheim, V. P. Clark, J. A. Turner, B. A. Mueller, V. Magnotta, et al. The mcic collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. *Neuroinformatics*, 11(3):367–388, 2013. 6

[12] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT Press, 2016. 5

[13] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017. 2

[14] A. J. Holmes, M. O. Hollinshead, T. M. OKeefe, V. I. Petrov, G. R. Fariello, L. L. Wald, B. Fischl, B. R. Rosen, R. W. Mair, J. L. Roffman, et al. Brain genomics superstruct project initial data release with structural, functional, and behavioral measures. *Scientific data*, 2, 2015. 6

[15] J. Immerkaer. Fast noise variance estimation. *Computer vision and image understanding*, 64(2):300–302, 1996. 4

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. 2

[17] T. Kapur, W. E. L. Grimson, W. M. Wells, and R. Kikinis. Segmentation of brain tissue from magnetic resonance images. *Medical image analysis*, 1(2):109–127, 1996. 2

[18] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 3, 4

[19] A. Klein and J. Tourville. 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in neuroscience*, 6:171, 2012. 5

[20] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007. 6

[21] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011. 6

[22] M. P. Milham, D. Fair, M. Mennes, S. H. Mostofsky, et al. The adhd-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in systems neuroscience*, 6:62, 2012. 6

[23] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 565–571. IEEE, 2016. 1

[24] P. Moeskops, M. Veta, M. W. Lafarge, K. A. Eppenhof, and J. P. Pluim. Adversarial training and dilated convolutions for brain mri segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 56–64. Springer, 2017. 2

[25] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett. Ways toward an early diagnosis in Alzheimers disease: the Alzheimers Disease Neuroimaging Initiative (ADNI). *Alzheimer's & Dementia*, 1(1):55–66, 2005. 6

[26] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, R. Guerrero, S. Cook, A. de Marvao, D. O'Regan, et al. Anatomically constrained neural networks (acnn): Application to cardiac image enhancement and segmentation. *arXiv preprint arXiv:1705.08302*, 2017. 2

[27] B. Patenaude, S. M. Smith, D. N. Kennedy, and M. Jenkinson. A Bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage*, 56(3):907–922, 2011. 2

[28] H. Ravishankar, R. Venkataramani, S. Thiruvenkadam, P. Sudhakar, and V. Vaidya. Learning and incorporating shape models for semantic segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 203–211. Springer, 2017. 2

[29] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 1, 2

[30] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 556–564. Springer, 2015. 2

[31] A. G. Roy, S. Conjeti, D. Sheet, A. Katouzian, N. Navab, and C. Wachinger. Error corrective boosting for learning fully convolutional networks with limited data. *arXiv preprint arXiv:1705.00938*, 2017. 1

[32] M. R. Sabuncu, B. T. Yeo, K. Van Leemput, B. Fischl, and P. Golland. A generative model for image segmentation based on label fusion. *IEEE transactions on medical imaging*, 29(10):1714–1729, 2010. 2

[33] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated model-based tissue classification of MR images of the brain. *IEEE transactions on medical imaging*, 18(10):897–908, 1999. 2

[34] C. Wachinger, M. Reuter, and T. Klein. Deepnat: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage*, 2017. 1, 2

[35] W. M. Wells, W. E. L. Grimson, R. Kikinis, and F. A. Jolesz. Adaptive segmentation of MRI data. *IEEE transactions on medical imaging*, 15(4):429–442, 1996. 2

[36] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012. 6

[37] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017. 2